Weekly Report

Yuxin Ma

08.20.2012 - 08.26.2012

Readings

KDD 2012 & SIGIR 2012 Some papers related to trust relation in social network came out in these conferences. It seems that the interpretation of data models has become a new research issue in data mining which visualization can be applied on. The papers are:

Transparent User Models for Personalization
eTrust: Understanding Trust Evolution in an Online World
and *Friend or Frenemy? Predicting Signed Ties in Social Networks*

In the next few weeks I will do more study on social trust relations.

Active learning and its applications The ideas related to active learning comes from *Research directions in data wrangling: Visualizations and transformations for usable and credible data*. In visual assessment and specification of automated methods, the author declares that it is a new research direction to apply visual-analysis-based method on proving data-processing algorithms. Moreover, he emphasizes that some approaches such as active learning can be applied in this process in order to not only visualize the result of automated algorithms but also transmit response from analysts to algorithms for guiding them into better performance, which is very similar to **active learning**.

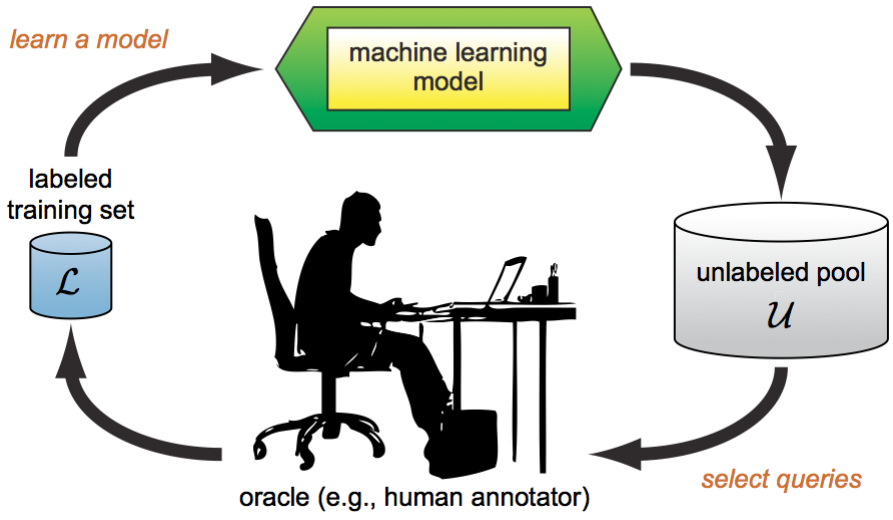

Figure 1: The labeled training set is generated by an oracle after asking queries by the active learner(machine learning model).

Active learning can be considered as an improvement of traditional machine learning-based algorithms to reduce training cost with nearly no loss of performance. The hypothesis in active learning scenario is that the learning algorithm is able to choose which it learns from training data. It is a solution for applications that is difficult to gain labels for training data(e.g. image tagging, which is a quite time-consuming task for analysts) by asking queries of unlabeled data which can be labeled by an "oracle" for the active learner. Figure 1 shows the training cycle of an active learning algorithm.

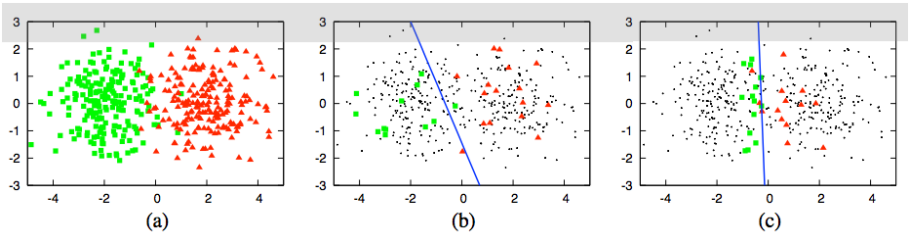

Figure 2: An active learning example.

Figure 2 is an illustrative example of active classifier. (a) is the ground-truth of 2 classes. For a traditional logistic regression model with 30 randomly labeled training instances, the result of classification is shown in (b). With active queries of "instances around the class boundary", the model will gain more accuracy as shown in (c).

There may be some ways of the combination of active learning method and visual analysis approach:

- From the aspect of active learning algorithms, there is a way of integrating visual analysis component right into the process of specified algorithm. Human annotators can access queries from the algorithm and find out answers with visual analysis approach on the data (or the algorithm itself can be visually accessed by analysts?) with much more interaction with the automated methods.
- In the framework of visual analytics the active data mining methods might be an extended part. (not very clear; need more thinking)
- (an application) The active learning methods can probably be applied in multidimensional projection in the situation that some constraints and query uncertainty exists in the dataset. There are some similar projects in clustering with active learning methods but not in multidimensional projection.

Practice & Skills

- **Heterogenous Computation in Data Visualization** During the summer vacation I found that in openFrameworks there are some OpenCL add-ons for high-performance rendering and data analysis. I will follow this issue continuously for it might be a solution for handling large-scale data visualization.
- **GraphLabAPI and GraphCHI** This library can be one of the best solutions I have ever met for graph applications. I have tested the PageRank toolkit from GraphLabAPI

which just takes less than 40 seconds on Slashdot dataset with ∼80,000 nodes and ∼500,000 edges, while in Gephi the time is more than 3 minutes. GraphLabAPI supports OpenMPI for parallelism as well.

Research

A performance issue from community constraints has been fixed. It may increase the speed of "accept all" operation. The next few days I will test this module and try to load the larger Gowalla dataset. Also in this week the comparing results of multiple community detection algorithms has been accomplished by Xu Jiayi.

Miscellaneous

- **Some Open Courses on Coursea.org** In coursea.org I found some courses related to big data, data science and social network analysis. I have signed up and enrolled in these courses which will begin in next month.

Plan for Next Week

- Test modified modules on larger dataset;
- finish images used in the textbook and review again.

References

[1] J. Tang, H. Gao, and H. Liu, "eTrust: Understanding Trust Evolution in an Online World," presented at the KDD '12: Proceedings of the eighteenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2012.

[2] K. El-Arini, U. Paquet, R. Herbrich, and J. Van Gael, "Transparent User Models for Personalization," presented at the KDD '12: Proceedings of the eighteenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2012.

[3] S. H. Yang, A. J. Smola, B. Long, and H. Zha, "Friend or frenemy?: predicting signed ties in social networks," presented at the SIGIR '12: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012.

[4] B. Settles, "Active learning literature survey," University of Wisconsin, Madison, 2010.

[5] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning," presented at the KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002.